UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/006,818	12/06/2001	Kevin P. Baker	GNE.2630P1C4	1321

7590 11/24/2004

Ginger R. Dreger
Knobbe Martens Olson & Bear
Sixteenth Floor
620 Newport Center Drive
Newport Beach, CA 92660

EXAMINER	
HAMUD, FOZIA M	
ART UNIT	**PAPER NUMBER**
1647	

DATE MAILED: 11/24/2004

Please find below and/or attached an Office communication concerning this application or proceeding.

Office Action Summary	Application No.	Applicant(s)
	10/006,818	BAKER ET AL.
	Examiner	Art Unit
	Fozia M Hamud	1647

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If the period for reply specified above is less than thirty (30) days, a reply within the statutory minimum of thirty (30) days will be considered timely.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on 19 August 2004.

2a)☒ This action is **FINAL**. 2b)☐ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) 28-32 is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) 28-32 is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☐ The drawing(s) filed on _____ is/are: a)☐ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)

2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)

3) ☒ Information Disclosure Statement(s) (PTO-1449 or PTO/SB/08) Paper No(s)/Mail Date 08/19/04.

4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____ .

5) ☐ Notice of Informal Patent Application (PTO-152)

6) ☐ Other: _____.

DETAILED ACTION

1. Receipt of Applicants' amendment and arguments, filed on 19 August 2004 is acknowledged. Claim 28 is currently amended and claim 33 is currently cancelled. Applicants' request to pursue any cancelled subject matter in subsequent continuation applications is noted.

Status of Claims:

1b. Claims 1-27 and 33 have been cancelled. Claims 28-32 are pending and under consideration.

1c. Receipt of Applicant's declaration under 37 C.F.R §1.132, filed by Dr. Avi Ashkenazi, Dr. Audrey Goddard and Dr. Paul Polakis filed on 19 August 2004 is also acknowledged.

2. ***Priority:***

Applicants submit that the results of the gene amplification assay disclosed in parent applications 60/162,506, filed 29 October 1999 priority for which has been claimed in the current application, provides a specific and substantial asserted utility for the claimed invention. Therefore, Applicants contend that the present application is entitled to the filing date of 29 October 1999.

This argument is not found persuasive. The claims of the instant invention are drawn to antibodies that bind to the polypeptide of SEQ ID NO:77. However, said subject matter is not supported by the disclosure in the international application 60/162,506, filed 29 October 1999, since said the prior application does not provide a specific and substantial asserted utility or a well established utility for the claimed

invention. As was previously stated and will be discussed in the following sections, the gene amplification assay described in the parent application does not provide a specific and substantial asserted utility for antibodies that bind to the polypeptide of SEQ ID NO:77, because the assay shows that DNA sequences encoding the polypeptide of SEQ ID NO:77 is amplified in the genome of certain human lung, colon and/or breast cancers and/or cell lines. However, the increased copy number of PRO1293 DNA in said tumors, does not provide a readily apparent use for antibodies that bind to the polypeptide of SEQ ID NO:77, because the assay does not show that the polypeptide is also amplified in these tumors.

Accordingly, the subject matter defined in claims 28-32 is afforded an effective filing date of 12/06/2001 which is the filing date of the current application.

Response to Applicants' arguments:

Claim Rejections under 35 U.S.C. §101/112:

35 U.S.C. 101 reads as follows:

> Whoever invents or discovers any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefor, subject to the conditions and requirements of this title.

The following is a quotation of the first paragraph of 35 U.S.C. 112:

> The specification shall contain a written description of the invention, and of the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the same and shall set forth the best mode contemplated by the inventor of carrying out his invention.

3. Claims 28-32 stand rejected under 35 U.S.C. §101, because the claimed invention is not supported by either a specific and substantial asserted utility or a well established utility, and are also rejected under 35 U.S.C. 112, first paragraph, for

reasons of record set forth in the office actions mailed on 19 March 2004. Specifically, since the claimed invention is not supported by either a credible, specific and substantial asserted utility or a well established utility for the reasons set forth above, one skilled in the art clearly would not know how to use the claimed invention.

3a. Applicant's arguments (submitted with the amendment of 19 August 2004) have been fully considered but are not found to be persuasive for the following reasons. The Ashkenazi, Dr. Goddard and Polakis declarations under 37 CFR 1 .132 filed 19 August 2004 are also insufficient to overcome the rejection of claims 28-32 based upon 35 U.S.C. §101 and 1 12, first paragraph as set forth in the last Office action for the following reasons.

3b. Applicants argue that the gene amplification is an essential mechanism for oncogene activation, and that this assay is well described in Example 143 of the present application. Applicants submit that there was a 2 to 8 fold increase of PRO1293 gene in lung and colon tumors. Applicants review Dr. Goddards' declaration, which states that the gene amplification technique used in the present specification is sensitive enough to detect a 2 fold increase in gene copy number in a tumor tissue compared to a normal tissue is significant and useful in a diagnostic manner.

This argument is fully considered, but is not found persuasive. It is not disputed that the gene amplification assay is useful in diagnostic manner and that the assay is well described in the instant specification, however, the instant specification does not demonstrate that the increased copy number of PRO1293 DNA in lung and colon tumors, leads to an increased expression of PR01293 polypeptide in these tumors.

Therefore, since Applicants do not provide information regarding the level of expression, an activity, or a role in cancer or any other disease for the PRO1293 polypeptide or the claimed antibody which binds to said polypeptide, both the polypeptide and the antibody lack a substantial utility or well established utility

3c. Applicants review the evidentiary standard regarding the legal presumption of utility. Applicant argues that the USPTO has not met its burden of overcoming the presumption of the truth of an asserted utility. This has been fully considered but is not found to be persuasive.

The examiner takes no issue with Applicant's discussion of the evidentiary standard regarding the legal presumption of utility. Furthermore, the rejection does not question the presumption of truth, or credibility, of the asserted utility.

The asserted utilities of cancer diagnostics for the claimed antibody that binds to the polypeptide of SEQ ID NO:77, are credible and specific. However, they are not substantial. The data set forth in the specification are preliminary at best. As the courts have discussed in Brenner v. Manson, 148 U.S.P.Q. 689 (Sup. Ct, 1966), an asserted utility must exist in currently available form. The specification indicates that the PRO1293 gene is amplified in certain cancers. However, the literature reports that gene amplification does not necessarily result in increased expression at the mRNA and polypeptide levels. For example, Hu et al. (2003, Journal of Proteome Research 2:405-412) analyzed 2286 genes that showed a greater than 1-fold difference in mean expression level between breast cancer samples and normal samples in a micoarray (p. 408, middle of right column). Hu et al. discovered that, for genes displaying a 5-fold

change or less in tumors compared to normal, there was no evidence of a correlation between altered gene expression and a known role in the disease. However, among genes with a 10- fold or more change in expression level, there was a strong and significant correlation between expression level and a published role in the disease (see discussion section).

3d. Applicant argues that even if a prima facie case of lack of utility has been established, it should be withdrawn on consideration of the totality of the evidence. Specifically, Applicant refers to the Ashkenazi declaration filed under 37 CFR § 1.132 with the amendment. The declaration and arguments assert that, even when amplification of a gene in a tumor does not correlate with an increase in polypeptide expression, the absence of the gene product over-expression still provides significant information for cancer diagnosis and treatment.

This has been fully considered but is not found to be persuasive. The examiner agrees that evidence regarding lack of over-expression would also be useful; unfortunately, there is no evidence as to whether the gene products (such as the polypeptide) are over-expressed or not. Further research is required to determine such. Thus, the asserted utility is not present in currently available form, and is not substantial. Applicant provides evidence in the form of a publication by Hanna et al., attached to the amendment. Applicant urges that the publication evidences that the HER-2/neu gene is over-expressed in breast cancers, and teaches that diagnosis of breast cancer includes testing both the amplification of the HER-2/neu gene as well as over-expression of the HER-2/neu gene product. Applicant argues that the disclosed

assay leads to a more accurate classification of the cancer and a more effective treatment of it. The examiner agrees. In fact, Hanna et al. supports the rejection, in that Hanna et al. show that gene amplification does not reliably correlate with polypeptide over-expression, and thus the level of polypeptide expression must be tested empirically. The specification does not provide this further information, and thus the skilled artisan must perform additional experiments. Since the asserted utility for the claimed polypeptides is not in currently available form, the asserted utility is not substantial.

3e. Applicant refers to three additional articles (Orntoft et al., Hyman et al. and Pollack et al.) as providing evidence that gene amplification generally results in elevated levels of the encoded polypeptide. Applicant characterizes Orntoft et al. as teaching in general (18 of 23 cases) chromosomal areas with more than 2-fo1d gain of DNA showed a corresponding increase in mRNA transcripts. Applicant characterizes Hyman et al. as providing evidence of a prominent global influence of copy number changes on gene expression levels. Applicant characterizes Pollack et al. as teaching that 62% of highly amplified genes show moderately or highly elevated expression and that, on average, a 2-fold change in DNA copy number is associated with a 1.5-fold change in mRNA levels.

This has been fully considered but is not found to be persuasive. Orntoft et al. appear to have looked at increased DNA content over large regions of chromosomes and comparing that to mRNA and polypeptide levels from the chromosomal region. Their approach to investigating gene copy number was termed CGH. Orntoft et al. do

not appear to look at gene amplification, mRNA levels and polypeptide levels from a single gene at a time. The instant specification reports data regarding amplification of individual genes, which may or may not be in a chromosomal region, which is highly amplified. Orntoft et al. concentrated on regions of chromosomes with strong gains of chromosomal material containing clusters of genes (p. 40). This analysis was not done for PRO1293 in the instant specification. That is, it is not clear whether or not PRO1293is in a gene cluster in a region of a chromosome that is highly amplified. Therefore, the relevance of Orntoft et al. is not clear. Hyman et al. used the same CGH approach in their research. Less than half (44%) of highly amplified genes showed mRNA over*expression (abstract). Polypeptide levels were not investigated. Therefore, Hyman et al. also do not support utility of the polypeptides of the instant invention. Pollack et al. also used CGH technology, concentrating on large chromosome regions showing high amplification (p. 12965). Pollack et al. did not investigate polypeptide levels. Therefore, Pollack et al. also do not support the asserted utility of the claimed invention. Importantly, none of the three papers reported that the research was relevant to identifying probes that can be used as cancer diagnostics. The three papers state that the research was relevant to the development of potential cancer therapeutics, but also clearly imply that much further research was needed before such therapeutics were in readily available form. Accordingly, the specification's assertions that antibodies that bind to PRO1293 polypeptides have utility in the fields of cancer diagnostics are not substantial.

3f. Applicant presents a declaration by Dr. Polakis filed with the response under 37 CFR 1.132. In the declaration, Dr. Polakis states that the primary focus of the Tumor Antigen Project was to identify tumor cell markers useful as targets for cancer diagnostics and therapeutics. Dr. Polakis states that approximately 200 gene transcripts were identified that are present in human tumor cells at significantly higher levels than in corresponding normal human cells. Dr. Polakis states that antibodies to approximately 30 of the tumor antigen polypeptides have been developed and used to show that approximately 80% of the samples show correlation between increased mRNA levels and changes in polypeptide levels. Dr. Polakis states that it remains a central dogma in molecular biology that increased mRNA levels are predictive of corresponding increased levels of the encoded polypeptide. Dr. Polakis characterizes the reports of instances where such a correlation does not exist as exceptions to the rule.

This has been fully considered but is not found to be persuasive. First, it is important to note that the instant specification provides no information regarding increased mRNA levels of PRO1293 in lung or colon cancer samples relevant to normal samples. Only gene amplification data was presented. Therefore, the declaration is insufficient to overcome the rejection of claims 28-32 based upon 35 U.S.C. §101 and §112, first paragraph, since it is limited to a discussion of data regarding the correlation of mRNA levels and polypeptide levels, and not gene amplification levels and polypeptide levels. Furthermore, the declaration does not provide data such that the examiner can independently draw conclusions. Only Dr. Polakis' conclusions are provided in the declaration. There is no evidentiary support to Dr. Polakis' statement that it remains a

central dogma in molecular biology that increased mRNA levels are predictive of corresponding increased levels of the encoded polypeptide. Finally, it is noted that the literature cautions researchers from drawing conclusions based on small changes in transcript expression levels between normal and cancerous tissue. (See, Hu et al., cited in paragraph 3c of this office action).

3g. Applicant presents a declaration by Dr. Goddard filed with the response under 37 CFR §1.132, however, this declaration is insufficient to overcome the rejection of claims 28-32 based upon 35 U.S.C §101/112.

The Declaration submitted by Dr. Goddard has been fully considered, but is deemed unpersuasive to overcome the rejection of claims 28-32 based upon 35 U.S.C 101/112. Dr. Goddard submits references that describe the gene amplification technique used in the present application and references that attest to the use of this technique in diagnostic and prognostic fashion. Finally, Dr. Goddard states that the gene amplification technique used in the present specification is sensitive enough to detect a 2 fold increase in gene copy number in a tumor tissue compared to a normal tissue is significant and useful in a diagnostic manner.

This argument is not found persuasive. Dr. Goddard's assertion that gene amplification is sensitive enough to detect a 2 fold increase and the a 2 fold increase in gene copy number in a tumor tissue compared to a normal tissue is significant and useful in a diagnostic manner, is correct. However, instant specification does not demonstrate that the increased copy number of PRO1293 DNA in lung and colon tumors, leads to an increased expression of PR01293 polypeptide in these tumors.

Therefore, since Applicants do not provide information regarding the level of expression, an activity, or a role in cancer or any other disease for the PRO1293 polypeptide or the claimed antibody which binds to said polypeptide, both the polypeptide and the antibody lack a substantial utility or well established utility.

For all of these reasons, the rejection claims 28-32 made under 35 U.S.C. §101 and §112 is maintained.

Claim Rejections - 35 U.S.C. §102:

The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

> A person shall be entitled to a patent unless –
>
> (a) the invention was known or used by others in this country, or patented or described in a printed publication in this or a foreign country, before the invention thereof by the applicant for a patent.

4a. Claims 28-32 stand rejected under U.S.C. § 102 (a) as being anticipated by Botstein et al (WO2000053751; published 14 September 2000).

Applicants submit that the current application is entitled to the filing date of 29 October 1999, (60/162,506). Hence, Applicants submit that WO2000053751 published on 14 September 2000 is not prior art under 102(a).

This argument is not found persuasive, because the invention of instant claims 28-32 are not entitled for the effective filing date of the priority application 29 October 1999, which is the filing date of 60/162,506, but is rather entitled to the filing date of the instant application, which is 12/06/2001, because the parent application does not teach how to use the claimed invention in a manner that satisfies the requirements, under 35 U.S.C. 112, first paragraph. See paragraph 3 of this office action.

Thus, WO2000053751 published on 14 September 2000 is prior art under 102(a).

Conclusion:

5. No claim is allowed.

THIS ACTION IS MADE FINAL. Applicant is reminded of the extension of time policy as set forth in 37 CFR 1.136(a).

A shortened statutory period for reply to this final action is set to expire THREE MONTHS from the mailing date of this action. In the event a first reply is filed within TWO MONTHS of the mailing date of this final action and the advisory action is not mailed until after the end of the THREE-MONTH shortened statutory period, then the shortened statutory period will expire on the date the advisory action is mailed, and any extension fee pursuant to 37 CFR 1.136(a) will be calculated from the mailing date of the advisory action. In no event, however, will the statutory period for reply expire later than SIX MONTHS from the mailing date of this final action.

Advisory Information:

Any inquiry concerning this communication or earlier communications from the examiner should be directed to Fozia M Hamud whose telephone number is (571) 272-0884. The examiner can normally be reached on Monday, Thursday-Friday, 6:00 am to 4:00 pm.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Brenda G Brumback can be reached on (571) 272-0961. The fax phone number for the organization where this application or proceeding is assigned is 703-872-9306.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free).

Fozia Hamud
Patent Examiner
Art Unit 1647
16 November 2004

JANET ANDRES
PRIMARY EXAMINER

08/19/04

INFORMATION DISCLOSURE STATEMENT PTO-1449	ATTY. DOCKET NO.: 39780-2630 P1C4	SERIAL NO.: 10/006,818
	APPLICANT : Kevin P. BAKER, et al.	
	FILING DATE: December 6, 2001	GROUP: 1647

U.S. PATENT DOCUMENTS

EXAMINER'S INITIALS	PATENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE

FOREIGN PATENT DOCUMENTS

EXAMINER'S INITIALS	PATENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	TRANSLATION YES	TRANSLATION NO
						☐	☐
						☐	☐
						☐	☐

OTHER DOCUMENTS (Including Author, Title, Date, Pertinent Pages, Etc.)

FH	Dayhoff, Accession No.: P_AAY99362, WO 200012708-A2, Pub Date: March 9, 2000, Baker, K., et al.
	Dayhoff, Accession No.: P_AAB24031, WO 200053750-A1, Pub Date: Sep. 14, 2000, Botstein, D., et al.
	Dayhoff, Accession No.: P_AAE06578, WO 200149728-A2, Pub Date: July 12, 2001, Kato, S., et al.
	Dayhoff, Accession No.: P_AAB36107, WO 200063438-A2, Pub Date: October 26, 2000, Gould-Rothberg, B.E., et al.
	Dayhoff, Accession No.: CAC51152.1, WO 0149728-A, Pub Date: July 12, 2001, Kato, S., et al.
	Dayhoff, Accession No.: AAH06213.1, Direct Submission, Submitted: April 9, 2001, Strausberg, R., et al.
	Dayhoff, Accession No.: AAH17312.1, Direct Submission, Submitted: November 5, 2001, Strausberg, R., et al.
	Dayhoff, Accession No.: NP_115724.1, Glia 44 (3), pp 190-204, 2003, Yonezawa, T., et al.
	Dayhoff, Accession No.: P_AAB92464, EP 1074617-A2, Pub Date: Feb. 7, 2001, Ota, T., et al.
	Dayhoff, Accession No.: BAB55010.1, Direct Submission, Submitted: May 10, 2001, Isogai, T., et al.
	GenBank, Accession No.: P_AAC58113, WO 200053750-A1, Pub Date: Sep. 14, 2000, Botstein, D., et al.
	GenBank, Accession No.: P_AAA37044, WO 200012708-A2, Pub Date: March 9, 2000, Baker, K., et al.
	GenBank, Accession No.: P_AAC76236, WO 200058473-A2, Pub Date: October 5, 2000, Shimkets, R.A., et al.

EXAMINER	[signature]	DATE CONSIDERED	11/12/04

EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609; draw line through citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

INFORMATION DISCLOSURE STATEMENT PTO-1449	ATTY. DOCKET NO.: 39780-2630 P1C4	SERIAL NO.: 10/006,818
	APPLICANT : Kevin P. BAKER, et al.	
	FILING DATE: December 6, 2001	GROUP: 1647

U.S. PATENT DOCUMENTS

EXAMINER'S INITIALS	PATENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE

FOREIGN PATENT DOCUMENTS

EXAMINER'S INITIALS	PATENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	TRANSLATION YES	TRANSLATION NO
						☐	☐
						☐	☐
						☐	☐

OTHER DOCUMENTS (Including Author, Title, Date, Pertinent Pages, Etc.)

PH	GenBank, Accession No.: BC006213, Direct Submission, Submitted: April 9, 2001, Strausberg, R., et al.
	GenBank, Accession No.: P_AAZ52368, WO 200021986-A2, Pub Date: April 20, 2000, Walker, M. G., et. al.
	GenBank, Accession No.: BC017312; Direct Submission, Submitted: November 5, 2001, Strausberg, R., et. al.
	GenBank, Accession No.: P_AAC67985, WO 200063438-A2, Pub Date: Oct. 26, 2000, Gould-Rothberg, B.E., et al.
	GenBank, Accession No.: P_AAD12573, WO 200149728-A2, Pub Date: July 12, 2001, Kato, S., et al.
	GenBank, Accession No.: AX191506.1, WO 0149728-A, Pub Date: July 12, 2001, Kato, S., et al.
	Hanna, J.S., et al., "HER-2/neu Breat Cancer Predictive Testing", Oathology Associates Medical Laboratores, August (1999).
	Hyman, Elizabeth, et al., "Impact of DNA Amplification on Gene Expression Patterns in Breast Cancer[1,2]", *Cancer Research* 62 6240-6245, November (2002).
	Orntoft, Torben F., et al., "Genome-wide Study of Gene Copy Numbers, Transcripts, and Protein Levels in Pairs of Non-Invasive and Invasive Human Transitional Cell Carcinomas", *Molecular & Cellular Proteomics* 1:37-45, (2002).
	Pollack, Jonathan R., et al., "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors", *PNAS* Vol.99, 20:12963-12968, October (2002).

EXAMINER [signature]	DATE CONSIDERED 10/12/04

EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609; draw line through citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

Notice of References Cited	Application/Control No. 10/006,818	Applicant(s)/Patent Under Reexamination BAKER ET AL.	
	Examiner Fozia M Hamud	Art Unit 1647	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	Classification
	A	US-			
	B	US-			
	C	US-			
	D	US-			
	E	US-			
	F	US-			
	G	US-			
	H	US-			
	I	US-			
	J	US-			
	K	US-			
	L	US-			
	M	US-			

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	Hu et al., 2003, Journal of Proteome Research 2:405-412.
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

Fozia Hamud 11/12/04

research articles Journal of proteome research

Analysis of Genomic and Proteomic Data Using Advanced Literature Mining

Yanhui Hu, Lisa M. Hines, Haifeng Weng, Dongmei Zuo, Miguel Rivera, Andrea Richardson, and Joshua LaBaer*

Institute of Proteomics, Harvard Medical School—BCMP, 240 Longwood Avenue, Boston, Massachusetts 02115

Received March 13, 2003

High-throughput technologies, such as proteomic screening and DNA micro-arrays, produce vast amounts of data requiring comprehensive analytical methods to decipher the biologically relevant results. One approach would be to manually search the biomedical literature; however, this would be an arduous task. We developed an automated literature-mining tool, termed MedGene, which comprehensively summarizes and estimates the relative strengths of all human gene–disease relationships in Medline. Using MedGene, we analyzed a novel micro-array expression dataset comparing breast cancer and normal breast tissue in the context of existing knowledge. We found no correlation between the strength of the literature association and the magnitude of the difference in expression level when considering changes as high as 5-fold; however, a significant correlation was observed ($r = 0.41$; $p = 0.05$) among genes showing an expression difference of 10-fold or more. Interestingly, this only held true for estrogen receptor (ER) positive tumors, not ER negative. MedGene identified a set of relatively understudied, yet highly expressed genes in ER negative tumors worthy of further examination.

Keywords: bioinformatics • micro-array • text mining • gene-disease association • breast cancer

Introduction

At its current pace, the accumulation of biomedical literature outpaces the ability of most researchers and clinicians to stay abreast of their own immediate fields, let alone cover a broader range of topics. For example, to follow a single disease, e.g., breast cancer, a researcher would have had to scan 130 different journals and read 27 papers per day in 1999.[1] This problem is accentuated with high-throughput technologies such as DNA micro-arrays and proteomics, which require the analysis of large datasets involving thousands of genes, many of which are unfamiliar to a particular researcher. In any microarray experiment, thousands of genes may demonstrate statistically significant expression changes, but only a fraction of these may be relevant to the study. The ability to interpret these datasets would be enhanced if they could be compared to a comprehensive summary of what is known about all genes. Thus, there is a need to summarize existing knowledge in a format that allows for the rapid analysis of associations between genes and diseases or other specific biological concepts.

One solution to this problem is to compile structured digital resources, such as the Breast Cancer Gene Database[1] and the Tumor Gene Database.[2] However, as these resources are hand-curated, the labor-intensive review process becomes a rate-limiting step in the growth of the database. As a result, these databases have a limited scale and the genes are not selected in a systematic fashion.

An alternative approach is automated text mining; a method which involves automated information extraction by searching documents for text strings and analyzing their frequency and context. This approach has been used successfully in several instances for biological applications. In most cases, it has been applied to extract information about the relationships or interactions that proteins or genes have with one another, in the literature or by functional annotation.[3–7] Thus far, few publication have applied text-mining to examine the global relationships between genes and diseases. Perez-Iratxeta et al. automatically examined the GO (Gene Ontology) annotation of genes and their predicted chromosomal locations in order to identify genes linked to inherited disorders.[8]

To obtain a more global understanding of disease development, it would be valuable to incorporate information regarding all possible gene-disease relationships, including biochemical, physiological, pharmacological, epidemiological, as well as genetic. This information would enable comprehensive comparisons between large experimental datasets and existing knowledge in the literature. This would accomplish two things. First, it would serve to validate experiments by demonstrating that known responses occur as predicted. Second, it would rapidly highlight which genes are corroborated by the literature and which genes are novel in a given context. We have utilized a computational approach to literature mining to produce a

* To whom correspondence should be addressed: jlabaer@hms.harvard.edu.

10.1021/pr0340227 CCC: $25.00

Published on Web 06/13/2003

comprehensive set of gene-disease relationships. In addition, we have developed a novel approach to assess the strength of each association based on the frequency of citation and co-citation. We applied this tool to help interpret the data from a large micro-array gene expression experiment comparing normal and cancerous breast tissue.

Methods

MedGene Database. MedGene is a relational database, storing disease and gene information from NCBI, text mining results, statistical scores, and hyperlinks to the primary literature. MedGene has a web-based user interface for users to query the database (http://hipseq.med.harvard.edu/MedGene/).

Text Mining Algorithms. MeSH files were downloaded from the MeSH web site at NLM (Nation Library of Medicine) (http://www.nlm.nih.gov/mesh/meshhome.html) and human disease categories were selected. LocusLink files were downloaded from the LocusLink web site at NCBI (http://www.ncbi.nih.gov/LocusLink/). Official/preferred gene symbol, official/preferred gene name, and gene alternative symbols and names, all relevant annotations and URLs for each LocusLink record, were collected. Gene search terms were used for literature searching and included all qualified gene names, gene symbols, and gene family terms. Primary gene keys, predominantly qualified gene family terms and gene official/preferred symbols, were used to index Medline records. If the official/preferred gene symbols did not meet the standards to be an index, then qualified gene official/preferred names were used. A local copy of Medline records (up to July, 2002) was pre-selected.

A JAVA module examined the MeSH terms and then indexed each Medline record with the appropriate disease terms. A separate JAVA module was used to examine the titles and abstracts for gene search terms and then to index the gene-related Medline records with the relevant primary gene key(s).

Statistical Methods. For every gene and disease pair, we counted records that were indexed for both gene and disease (double positive hits), for disease only (disease single hits), for gene only (gene single hits), and for neither gene nor disease (double negative hits) to generate a 2×2 contingency table. On the basis of the contingency table-framework, we applied different statistical methods to estimate the strength of gene-disease relationships and evaluated the results. These methods included chi-square analysis, Fisher's exact probabilities, relative risk of gene, and relative risk of disease[16] (http://hipseq.med.harvard.edu/MedGene/). In addition, we computed the "product of frequency", which is the product of the proportion of disease/gene double hits to disease single hits and the proportion of disease/gene double hits to gene single hits. To obtain a normal distribution, we transformed all the statistical scores using the natural logarithm. We selected the log of the product of frequency (LPF) to validate MedGene and to use for the analysis with the micro-array data. Spearman rank-correlation coefficients were used to assess the linear relationship between LPF and micro-array fold change in expression level.

Global Analysis. Diseases with at least 50 related genes were selected for clustering analysis, and the LPF scores were normalized with total score for each disease. Hierarchical clustering was done with the "Cluster" software and the clustering result was visualized using "TreeViewer" (http://rana.lbl.gov/EisenSoftware.htm).

Breast Tissue Micro-Arrays. Eighty-nine breast cancer samples (79% ER-positive) and 7 normal breast tissue samples were selected from the Harvard Breast SPORE frozen tissue repository and were representative of the spectrum of histological types, grades, and hormone receptor immuno-phenotypes of breast cancer. Biotinylated cRNA, generated from the total RNA extracted from the bulk tumor, was hybridized to Affymetrix U95A oligo-nucleotide micro-arrays. These micro-arrays consist of 12 400 probes, which represent approximately 9000 genes. Raw expression values were obtained using GENE-CHIP software from Affymetrix, and then further analyzed using the DNA-Chip Analyzer (dChip) custom software.

Results

Automated Indexing of Medline Records by Disease and Gene. To study the gene-disease associations in the literature, we first compiled complete lists for human diseases and human genes. To index all Medline records that were relevant to human diseases, the Medical Subject Heading (MeSH) index of Medline records was utilized. MeSH is a controlled medical vocabulary from the National Library of Medicine and consists of a set of terms or subject headings that are arranged in both an alphabetic and an hierarchical structure. Medline records are reviewed manually and MeSH terms are added to each with software assistance.[9,10] Twenty-three human disease category headings along with all of their child terms (see the Supporting Information, Supplemental Table 1, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Table 1.html) were selected from the 2002 MeSH index creating a list of 4033 human diseases.

No index comparable to the MeSH index exists for genes, and thus, it was necessary to apply a string search algorithm for gene names or symbols found in Medline text. A complete list of genes, gene names, gene symbols, and frequently used synonyms were collected from the LocusLink database at NCBI,[11,12] which contains 53 259 independent records keyed by an official gene symbol or name (June 18th, 2002). For the purposes of this study, no distinction was made between genes and their gene products. Authors often use the same name for both, differentiating the two only by the use of italics, if at all. For the intended use of this study, this lack of distinction is unlikely to have a large effect and may in fact be beneficial.

Initial attempts to search the literature using these lists revealed several sources of false positives and false negatives (Table 1). False positives primarily arose when the searched term had other meanings, whereas false negatives arose from syntax discrepancies necessitating the development of filters to reduce these errors. The syntax issues were readily handled by including alternate syntax forms in the search terms. The false positive cases, caused by duplicative and unrelated meanings for the terms, were more difficult to manage. Where possible, case sensitive string mapping reduced inappropriate citations. In many cases, however, this was not sufficient and the terms had to be eliminated entirely, thereby reducing the false positive rate but unavoidably under-representing some genes.

For the purposes of data tracking, a primary gene key was selected to represent all synonyms that correspond to each gene. Medline records were indexed with a primary gene key when any synonym for that key was found in the title or abstract. Case-insensitive string mapping was used for all searches except as noted above. No additional weight was

Table 1. Systematic Sources of False Positives and False Negatives in Unfiltered Data[a]

source of error	error type	example	filter solution
gene symbol/name is not unique	false positive	*MAC*—myelin associated glycoprotein *MAC*—malignancy-associated protein	eliminate this term
gene symbol is unrelated abbreviation	false positive	*PA*—pallid homologue (mouse), pallidin (also abbrev. for Pennsylvania)	eliminate this term
gene symbol/name has language meaning	false positive	*WAS*—Wiskott–Aldrich Syndrome (also the word "was")	case-sensitive string search
nonstandard syntax	false negative	*BAG-1* instead of *BAG1*	add dash term
unofficial gene name/symbol	false negative	*P53* instead of *TP53*	add all gene nicknames
nonspecified gene name	false negative	estrogen receptor instead of Estrogen receptor 1	add family stem term

[a] In preliminary studies, Medline was searched for co-occurrence of genes and diseases and the resulting output was evaluated to identify error sources that were amenable to global filters. Each error source is categorized by the type of error it causes: false positives are suggested relationships that are not real and false negatives are real relationships that are underrepresented. The filter solutions used are indicated. Note that in some cases, the filter solution itself introduces error. In general, error rates maximized sensitivity, even at the expense of specificity if needed.

added for multiple occurrences of a term or the co-occurrence of multiple synonyms for the same gene key.

Medline records were searched with all qualified gene identifiers, such as the official/preferred gene symbol, the official/preferred gene name, all gene nicknames and all syntax variants. In situations where there are several members of a gene family or splice variants, some authors prefer to use a shortened gene family name, e.g., estrogen receptor instead of estrogen receptor 1 (*ESR1*), creating a source of false negatives. For this reason, gene family stem terms were created for all genes that have an alpha or numerical suffix (e.g., *IL2RA*, *TGFβ*, *ESR1*, etc.) and then used to search the literature. The family stem terms were handled separately from the specific gene names so that it would be clear when linkages were made to the gene family versus a specific member in that family.

To improve performance and accuracy, some pre-selection was applied to the records that were scanned. First, review articles were eliminated to avoid redundant treatment of citations. Second, non-English journals were removed because the natural language filters were only relevant to English publications. Finally, journals unlikely to contain primary data about gene-disease relationships were also removed (e.g., *Int. J. Health Educ.*, *Bedside Nurse*, and *J. Health Econ.*). Together, these filters reduced the 12 198 221 Medline publications (July 2002) by 37%.

Ranking the Relative Strengths of Gene-Disease Associations. In total, there were 618 708 gene-disease co-citations, in which 16% (8297) of all studied genes had been associated to a disease and 96% (3875) of all diseases had been associated to at least one gene. To rank the relative strengths of gene disease relationships, we tested several different statistical methods and examined the results. With the exception of the relative risk estimates, the methods provided similar results with respect to the rank order of the gene-disease association strengths. However, after comparing the results to other databases and after consulting disease experts, the log of the product of frequency (LPF) was selected for further analysis because it gave the best results overall.

Validation of MedGene. In developing this tool, it was important to minimize the number of missed genes (false negatives) and miscalled genes (false positives). However, in situations when these goals were in conflict, inclusiveness was prioritized. To determine the false negative rate in MedGene, breast cancer was used as a test case because it was associated with more genes than any other human disease and because there were several public databases that link genes to breast cancer. We compared the list of breast cancer-related genes from MedGene to these databases, illustrated in Figure 1. Among the 285 distinct breast cancer-related genes that were supported by at least one literature citation in these hand-curated databases, 26 were absent from MedGene, suggesting a false negative rate of approximately 9%. To determine why these were missed, all literature references for these genes (80

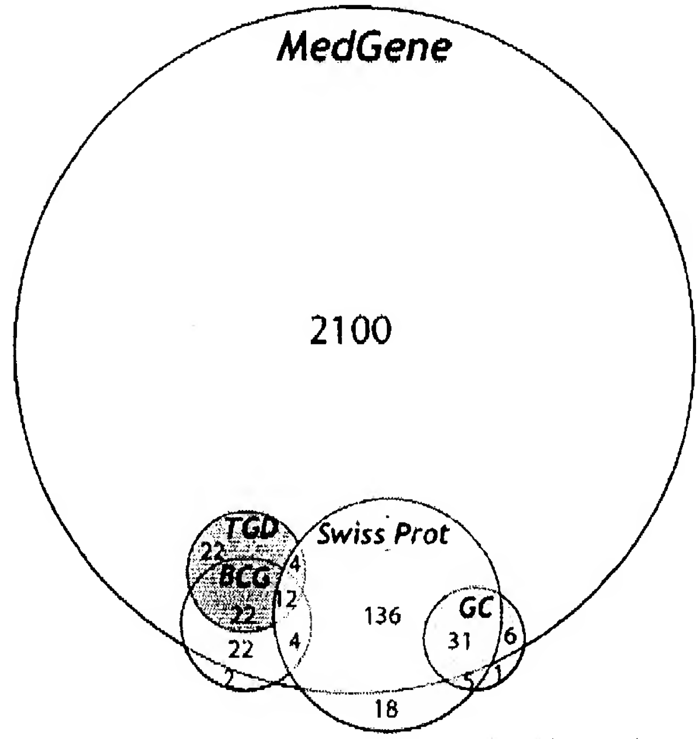

Figure 1. Estimation of the false negative rate by comparison with hand-curated databases. The breast cancer-related genes identified by MedGene were compared with those listed in several other databases including the Tumor Gene Database (TGD),[7] the Breast Cancer Gene Database(BCG),[1] GeneCards (GC)[17] and Swissprot.[18] Genes were considered false negatives if they were represented in at least one of these other databases and not in MedGene and their link to breast cancer was supported by at least one literature reference. All literature references were verified by manual review to confirm their validity. The number of genes in each database or shared by more than one database is indicated. The false negative rate was calculated by genes missed at MedGene (26)/total number of nonoverlapping genes in other databases (285).

papers) were reviewed manually (see the Supporting Information, Supplemental Table 2, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Table 2.html). Among these papers, most false negatives were caused by nonstandard gene terms or gene terms eliminated by our specificity filters. Few genes were missed because they were only mentioned in review papers (0.4%) or they appeared only in the body of the manuscript but not the abstract or title (1.1%). Of note, MedGene identified approximately 2000 additional breast cancer-related genes not listed in any other database.

To assess the false positive error rate, two complementary approaches were used: a detailed analysis of one disease and a global examination of 1000 diseases. The detailed approach examined the false positive error rate and its sources, whereas the global approach tested whether the overall results made biomedical sense.

Using the LPF, 1467 genes related to prostate cancer were assembled in rank order. We then retrieved approximately 300 Medline records each for the highest ranked 100 and the lowest ranked 200 genes and manually reviewed the titles and abstracts to determine the verity of the association. Nearly 80% of the highest ranked 100 genes fell into one of the five categories that reflect meaningful gene-disease relationships (see the Supporting Information, Supplemental Table 3, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Table 3.html). Among the lowest ranked 200 genes, approximately 70% reflected true relationships. Of the 600 records reviewed, there were only two in which the association between the gene and the disease was described as negative. Both were genes with very low scores. In both cases, the authors did not argue the absence of *any* relationship, but rather that a particular feature of the gene or protein was not shown to be related to human prostate cancer.[13,14]

The coincidence of some gene symbols with medical abbreviations, chemical abbreviations and biological abbreviations resulted in most of the false positives (see the Supporting Information, Supplemental Table 4, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Table 4.html), emphasizing the importance of the filters that were added in the search algorithm (Table 1). Without the filters, the false positive rate more than doubled, and the false negative rate rose dramatically (data not shown). For example, among the papers about breast cancer, there were only 12 Medline records that referred to *ESR1* and 10 to *ESR2*, whereas almost 2000 papers mentioned estrogen receptor without specifying *ESR1* or *ESR2*; this latter group was detected by the family stem term filter.

To further validate these results, a global analysis of the gene-disease relationships described by MedGene was performed. For this experiment, it was reasoned that the more closely related the diseases are to one another, the more they will be related to the same gene sets. Thus, if the relationships defined by MedGene accurately reflected the literature, then an unsupervised hierarchical clustering of the gene data should group diseases in a manner consistent with common medical thinking. Conversely, if the clustered diseases do not make sense biologically or medically, it may reflect excessive false positives, false negatives, or inappropriate scoring of the data.

To execute this experiment, the gene sets and the corresponding LPF values for 1000 randomly selected diseases (each with at least 50 gene relationships) were used as a dataset for clustering the diseases. A review of the results showed that the resulting disease clusters were indeed logical based upon common medical knowledge (see the Supporting Information, Supplemental Figure 1, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Figure 1.html). For example, in one such cluster shown in Figure 2, diabetes and its complications grouped together and were also closely linked to diseases associated with starvation states.

The number of genes associated with a given disease can be estimated by adjusting the MedGene number up by the false negative rate (~9%) and down by the false positive rate (~26% on average). Using this, the average disease has 103.7 ± 45.3 (mean ± s.d.) genes associated with it, although the range is quite broad with 2359 genes related to breast cancer, 2122 genes related to lung cancer and no genes related to a number of diseases.

Applying MedGene to the Analysis of Large Datasets. Access to a comprehensive summary of the genes linked to human diseases provided an opportunity to analyze data obtained from a high-throughput experiment. We compared the MedGene breast cancer gene list to a gene expression data set generated from a micro-array analysis comparing breast cancer and normal breast tissue samples. Micro-array analysis identified 2286 genes that had greater than a 1-fold difference in mean expression level between breast cancer samples and normal breast samples. Using MedGene, we sorted the 2286 genes into four classes: 555 genes directly linked to breast cancer in the literature by gene term search (first-degree association by gene name); 328 genes directly linked by family term search (first-degree association by family term); 1021 genes linked to breast cancer only through other breast cancer genes (second-degree association); and 505 genes not previously associated with breast cancer. (See the Supporting Information, Supplemental Figure 2, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Figure 2.html.) Among the 505 previously unrelated genes, 467 were either newly identified genes or genes that had not previously been associated with any disease. Among the remaining 38 genes, 9 had been related to other cancers, specifically esophageal, colon, uterine, skin, and cervix.

To determine whether the genes highlighted by the micro-array analysis were more likely to have been previously linked to breast cancer in the literature, we created a two-dimensional plot of the fold change of expression level between breast cancer and normal tissue versus the literature score (LPF) (Figure 3A). There was a broad spread of expression changes among the genes directly linked to breast cancer ranging from less than 1-fold change (68%) to over 40-fold (0.3%). Notably, the majority of genes with greater than 10-fold expression changes were linked to breast cancer by first-degree association.

Among all 754 genes directly linked to breast cancer in the literature, there was no correlation between LPF and micro array fold change ($r = 0.018$, p-value $= 0.62$). However, when we stratified the analysis based on the magnitude of the fold change, we observed an increasing trend in correlation (Figure 3B) suggesting that genes with a more substantial change in expression level were more likely to have a stronger association in the literature. For genes that had 10 fold change or more in expression level, the correlation increased to 0.41 (p-value $=$ 0.05).

When we evaluated the micro-array data separately for ER positive and ER negative tumors, the trend in correlation between fold change and literature score was highly dependent on estrogen receptor status. Interestingly, there was a similar trend in correlation for ER positive tumors, but no trend in correlation for ER negative tumors.

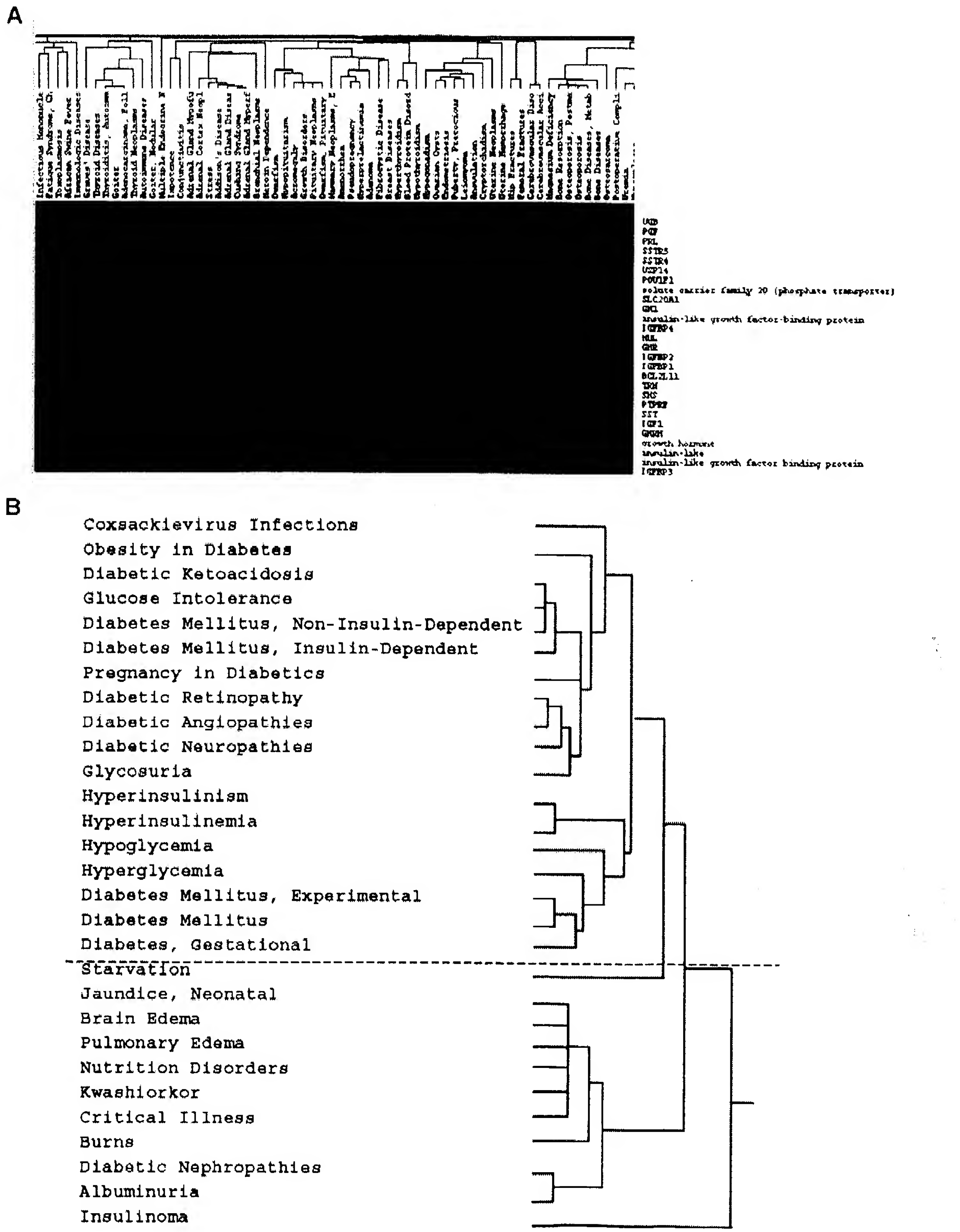

Figure 2. Global validation by clustering analysis. 2(**A**). The gene sets and the corresponding LPF values for 1000 diseases, each with at least 50 gene relationships, were used in an unsupervised clustering of the diseases based on the gene patterns associated with them. A sample of the data is shown here. 2(**B**). One of the resulting clusters is shown that corresponds to blood sugar states. Diabetes terms (above the line) and starvation states terms (under the line) clustered together. Within these groups, there is also clustering of diabetic small vessel complications, altered serum chemistries, nutritional disorders, etc.(Supplemental Figure 1: http://hipseq.med.harvard.edu/MedGene/publication/s_Figure 1.html).

Finally, to validate our findings, we computed similar correlations between the breast cancer expression data and LPF scores generated by MedGene for hypertension, a disease unrelated to breast cancer. As expected, we did not observe an increasing trend in correlation for hypertension.

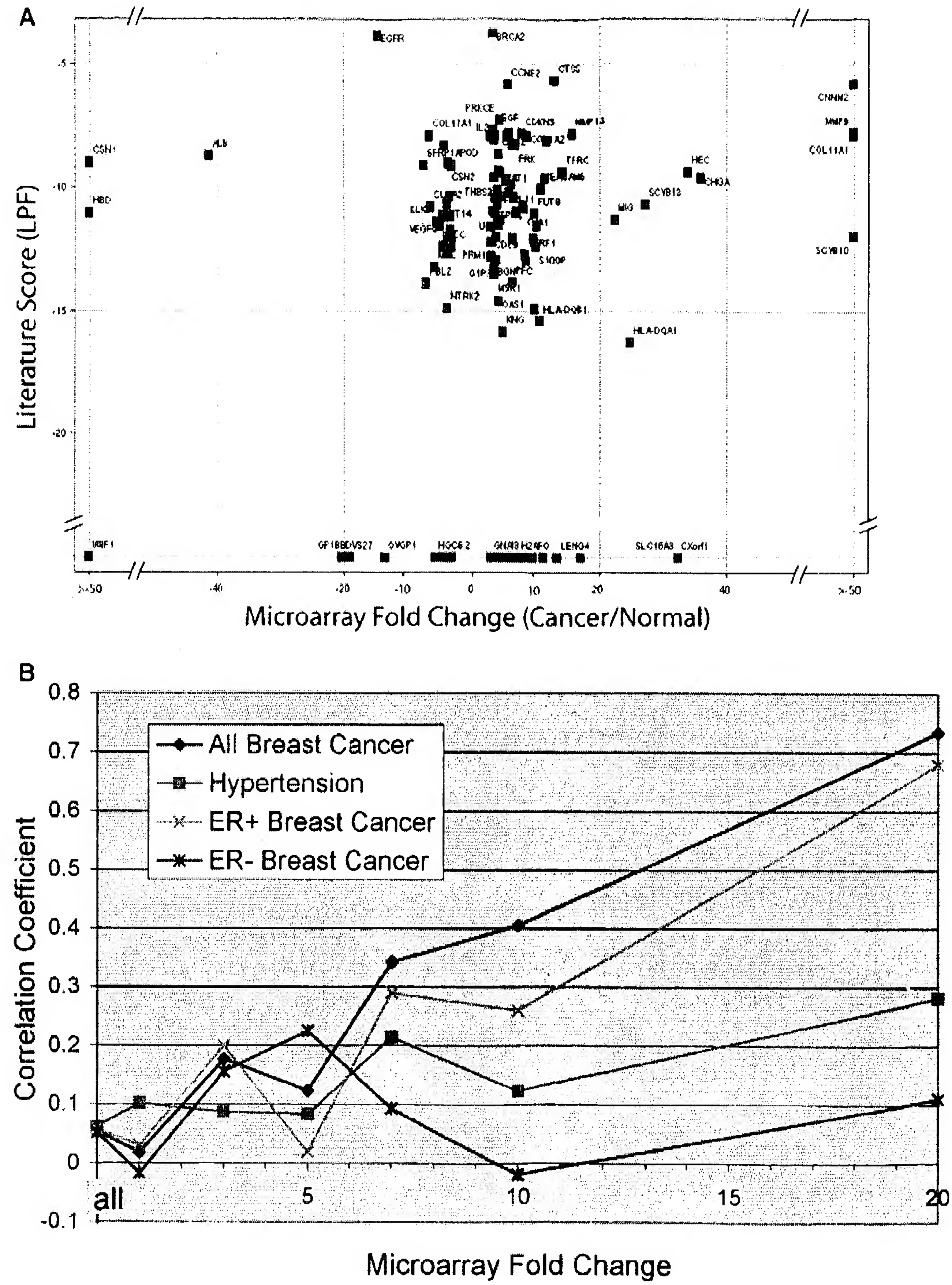

Figure 3. Relationship between literature score and functional data for breast cancer. **3A**. The data from an expression analysis of samples for breast tumors and normal breast tissue were analyzed to indicate the fold difference of expression level between breast tumor and normal sample (cutoff ≥ 3-fold change). The fold changes were plotted against the literature score for the same gene set. Green dots represent first-degree association by gene search, blue dots represent first-degree association by family search and red dots represent no-association. Some well-studied genes, such as BRCA2 (pink circle), are not reflected by a substantial difference in expression level. Furthermore, the majority of genes that have no association with breast cancer in the literature had less than 10-fold expression changes (shaded area). **3B**. The Spearman rank-correlation coefficients between literature score (LPF) and the fold change of expression level between tumor and normal breast samples (*y*-axis) in relation to the amount of fold change of expression level (*x*-axis). Gene rank lists were generated for breast cancer (blue) and hypertension (pink). Correlations were also computed between the breast cancer gene LPF scores and fold change expression data among estrogen receptor positive tumors only (light blue) and estrogen receptor negative tumors only (purple).

Table 2. Top 25 Genes Related to Selected Human Diseases[a]

breast neoplasms	hypertension	rheumatoid arthritis	bipolar disorder	atherosclerosis
estrogen receptor	*REN*	*RA*	*ERDA1*	apolipoprotein
PGR	*DBP*	*TNFRSF10A*	*SNAP29*	*APOE*
ERBB2	*LEP*	*CRP*	*PFKL*	*LDLR*
BRCA1	*AGT*	*AS*	*DRD2*	*ELN*
BRCA2	*INS*	*ESR1*	*TRH*	*ARG1*
EGFR	kallikrein	*HLA-DRB1*	*IMPA2*	*APOB*
CYP19	*ACE*	*DR1*	*HTR3A*	*APOA1*
TFF1	endothelin	interleukin	*DRD3*	*MSR1*
PSEN2	*S100A6*	*TNF*	*REM*	*LPL*
TP53	*BDK*	*IL6*	*KCNN3*	*PON1*
CES3	*DIANPH*	collagen	*DRD4*	plasminogen activator inhibitor
CEACAM5	*SAR1*	*IL1A*	*HTR2C*	*PLG*
ERBB3	*PIH*	*ACR*	*RELN*	vascular cell adhesion molecule
cyclin	*CD59*	*TNFRSF12*	*DBH*	*ATOH1*
COX5A	*ALB*	*IL2*	*MAOA*	*VWF*
cathepsin	*CYP11B2*	*CHI3L1*	*COMT*	*INS*
ERBB4	*MAT2B*	*IL8*	*HTR2A*	*ARG2*
TRAM	angiotensin receptor	interleukin 1	*SYNJ1*	*ABCA1*
CCND1	*AGTR2*	matrix metalloproteinase	*INPP1*	*OLR1*
EGF	*NPPA*	interferon	*NEDD4L*	collagen
MUC1	*LVM*	*CD68*	*FRA13C*	*MCP*
insulin-like	*DBH*	*IL4*	transducer of *ERBB2*	lipoprotein
BCL2	*NPY*	*IL17*	*BAIAP3*	*APOA2*
mucin	*POMC*	*MMP3*	*ATP1B3*	intercellular adhesion molecule
FGF3	neuropeptide	*SIL*	*DRD5*	*RAB27A*

[a] MedGene results for the top 25 genes associated with breast neoplasms, hypertension, rheumatoid arthritis, bipolar disorder, and atherosclerosis, respectively, ranked by LPF scores. The hyperlink to all the papers co-citing the gene and the disease is available at MedGene website (http://hipseq.med.harvard.edu/MedGene/).

Discussion

The Human Genome Project heralded a new era in biological research where the emphasis on understanding specific pathways has expanded to global studies of genomic organization and biological systems. High-throughput technologies can provide novel insight into comprehensive biological function but also introduces new challenges. The utility of these technologies is limited to the ability to generate, analyze, and interpret large gene lists. MedGene, a relational database derived by mining the information in Medline, was created to address this need. MedGene users can query for a rank-ordered list of human gene-disease relationships (Table 2) for one or more diseases. Each entry is hyperlinked to the original papers supporting each association and to other relevant databases.

MedGene is an innovative extension of previous text mining approaches. Perez-Iratxeta et al. used the GO annotation and their chromosomal locations to predict genes that may contribute to inherited disorders.[8] MedGene takes a broader view and includes all diseases and all possible gene-disease relationships. Furthermore, MedGene utilizes co-citation to indicate a relationship rather than GO annotation, which is limited to the subset of genes that have GO annotation. Our approach is complementary to that taken by Chaussabel and Sher, who used the frequency of co-cited terms to cluster genes into a hierarchy of gene-gene relationships.[6]

A unique aspect of this tool is the ability to assess the relative strengths of gene-disease relationships based on the frequency of both co-citation and single citation. This presupposes that most co-citations describe a positive association, often referred to as publication bias[15] and is supported by our observations that negative associations are rare (Supplemental Table 3: http://hipseq.med.harvard.edu/MedGene/publication/s_Table 3.html). Of course, relationships established by frequency of co-citation do not necessarily represent a true biological link; however, it is strong evidence to support a true relationship.

Another important feature of MedGene is the implementation of software filters that substantially reduced the error rate. We estimate that less than 10% of all associations were missed and at least 70% of even the weakest associations were real. For this study, all of the filters that we applied were general ones, e.g., expanding the list of all gene names to address the different syntax forms used by different journals, eliminating gene names that correspond to common English words, etc. The majority of the remaining search term ambiguities were idiosyncratic and difficult to identify systematically without causing a significant rise in false negatives. Alternative approaches, such as the examination of the nearest neighbor terms, need to be considered to further reduce the false positive rate.

It is not uncommon to see expression changes in microarray experiments as small as 2-fold reported in the literature. Even when these expression changes are statistically significant, it is not always clear if they are biologically meaningful. When comparing expression levels of disease to normal tissue, one expects an enrichment of known disease-related genes to appear in the altered expression group. MedGene provided a unique opportunity to test this notion in the context of existing knowledge on a novel breast cancer micro-array dataset. For genes displaying a 5-fold change or less in tumors compared to normal, there was no evidence of a correlation between altered gene expression and a known role in the disease. This

Table 3. Genes with Large Expression Changes in ER− but Not in ER+ Breast Tumors

gene symbol	fold change (ER+)	fold change (ER−)
KRTHB1	1.0	610.8
BRS3	1.2	89.4
DKK1	1.2	69.8
ZIC1	1.9	59.6
TLR1	1.0	38.5
KIAA0680	2.6	33.2
CDKN3	1.0	30.6
EBI2	4.0	27.9
GZMB	3.8	21.9
STK18	4.7	18.6
GPR49	1.0	14.6
MYO10	1.6	14.4
LAD1	−1.0	13.5
POLE2	4.2	13.0
HMG4	4.4	12.9
BCL2L11	−1.2	12.3
LRP8	2.9	12.2
CCNB2	1.0	11.8
CCNE2	4.0	11.6
FCB	−4.3	11.1
KNSL6	2.9	10.9
HIF5	3.0	10.2
SERPINH2	4.6	10.2
YAP1	1.0	10.0
LPHB	−1.3	−10.4
TCEA2	−1.1	−10.8
TFF1	1.3	−11.4
COL17A1	−4.1	−15.7
POP5	1.1	−16.2
BPAG1	−4.6	−22.3
PDZK1	−1.1	−36.8
VEGFC	−2.8	−51.5
MUC6	−1.4	−64.9
SERPINA5	−1.0	−83.1
MEIS1	−1.6	−85.9
CA12	2.4	−150.3

Table 3. MedGene identified a set of relatively understudied, yet highly expressed genes in ER negative, but not ER positive breast tumors. All of these genes have either never been co-cited with breast cancer or have a weak association except those marked with an *.

reflects the many genes whose role in breast cancer may not involve large changes in expression in sporadic tumors (e.g., *BRCA1* and *BRCA2*) and genes whose modest changes in expression may be unrelated to the disease. Strikingly, among genes with a 10-fold change or more in expression level, there was a strong and significant correlation between expression level and a published role in the disease, providing the first global validation of the micro-array approach to identifying disease-specific genes.

The results derived from MedGene have two implications. First, a careful hunt for corroborating evidence of a role in breast cancer should precede any further study of genes with less than 5-fold expression level changes. Second, any genes with 10-fold changes or more are likely to be related to breast cancer and warrant attention. It is likely that this threshold will change depending on the disease as well as the experiment.

Interestingly, the observed correlation was only found among ER-positive tumors, not ER-negative. This may reflect a bias in the literature to study the more prevalent type of tumor in the population. Furthermore, this emphasizes that caution must be taken when interpreting experiments that may contain subpopulations that behave very differently. The MedGene approach identified a set of relatively understudied, yet highly expressed genes in ER-negative tumors that are worthy of further examination (Table 3).

In conclusion, we have developed an automated method of summarizing and organizing the vast biomedical literature. To our knowledge, the resulting database is the most comprehensive and accurate of its kind. By generating a score that reflects the strength of the association, it provides an important tool for the rapid and flexible analysis of large datasets from various high-throughput screening experiments. Furthermore, it can be used for selecting subsets of genes for functional studies, for building disease-specific arrays, for looking at genes common to multiple diseases and various other high-throughput applications. In the future, it will be possible to enhance the utility of the MedGene database by building links between genes and other MeSH terms as well as other biological processes and concepts, such as cell division and responses to small molecules.

Acknowledgment. We would like to thank P. Braun, L. Garraway, J. Pearlberg, and other members of our institute for helpful discussion. Many thanks to the NLM (National Library of Medicine) for licensing of MEDLINE and the annotation effort of adding MeSH indexes for MEDLINE abstracts. This work was funded by grants from the Breast Cancer Research Foundation and an NHLBI PGA Grant (Vol HL66582-02).

Supporting Information Available: Twenty-three human disease category headings along with all of their child terms selected from the 2002 MeSH index (Supplemental Table 1); analysis of the causes of false negatives in MedGene (Supplemental Table); meaningful gene-disease relationships found in MedGene (Supplemental Table 3); causes for incorrect assignment of gene indexes (Supplemental Table 4); a review of the results, showing that the resulting disease clusters were indeed logical (Supplemental Figure 1); and a review of the results showing that among the 505 previously unrelated genes, 467 were either newly identified genes or genes that had not previously been associated with any disease (Supplemental Figure 2). This material is available free of charge via the Internet at http://pubs.acs.org and at the web sites mentioned in the text.

References

(1) Baasiri, R. A.; Glasser, S. R.; Steffen, D. L.; Wheeler, D. A. *Oncogene* **1999**, *18*, 7958–7965.
(2) Steffen, D. L.; Levine, A. E.; Yarus, S.; Baasiri, R. A.; Wheeler, D. A. *Bioinformatics* **2000**, *16*, 639–649.
(3) Marcotte, E. M.; Xenarios, I.; Eisenberg, D. *Bioinformatics* **2001**, *17*, 359–363.
(4) Ono, T.; Hishigaki, H.; Tanigami, A.; Takagi, T. *Bioinformatics* **2001**, *17*, 155–161.
(5) Jenssen, T. K.; Laegreid, A.; Komorowski, J.; Hovig, E. *Nat. Genet.* **2001**, *28*, 21–28.
(6) Chaussabel, D.; Sher, A. *Genome Biol.* **2002**, *3*, RESEARCH0055.
(7) Gibbons, F. D.; Roth, F. P. *Genome Res.* **2002**, *12*, 1574–1581.
(8) Perez-Iratxeta, C.; Bork, P.; Andrade, M. A. *Nat. Genet.* **2002**, *31*, 316–319.
(9) Funk, M. E.; Reid, C. A. *Bull. Med. Libr. Assoc.* **1983**, *71*, 176–183.
(10) Humphrey, S. M.; Miller, N. E. *J. Am. Soc. Inf. Sci.* **1987**, *38*, 184–196.
(11) Maglott, D. R.; Katz, K. S.; Sicotte, H.; Pruitt, K. D. *Nucleic Acids Res.* **2000**, *28*, 126–128.
(12) Pruitt, K. D.; Maglott, D. R. *Nucleic Acids Res.* **2001**, *29*, 137–140.
(13) Wadelius, M.; Andersson, A. O.; Johansson, J. E.; Wadelius, C.; Rane, E. *Pharmacogenetics* **1999**, *9*, 333–340.
(14) Adam, R. M.; Borer, J. G.; Williams, J.; Eastham, J. A.; Loughlin, K. R.; Freeman, M. R. *Endocrinology* **1999**, *140*, 5866–5875.
(15) Montori, V. M.; Smieja, M.; Guyatt, G. H. *Mayo Clin. Proc.* **2000**, *75*, 1284–1288.
(16) Denenberg, V. H. *Statistics Experimental Design for Behavioral and Biological Researchers*; Wiley-Liss: New York, 1976.
(17) Rebhan, M.; Chalifa-Caspi, V.; Prilusky, J.; Lancet, D. *Trends Genet.* **1997**, *13*, 163.
(18) Bairoch, A.; Apweiler, R. *Nucleic Acids Res.* **2000**, *28*, 45–48.

PR0340227

Organization TC 1600 Bldg./Room REMSEN
U. S. DEPARTMENT OF COMMERCE
COMMISSIONER FOR PATENTS
P.O. BOX 1450
ALEXANDRIA, VA 22313-1450
IF UNDELIVERABLE RETURN IN TEN DAYS

OFFICIAL BUSINESS

AN EQUAL OPPORTUNITY EMPLOYER

RETURNED TO SENDER
FORWARDING ORDER EXPIRED

RECEIVED

DEC 0 6 2004

TECH CENTER 1600/2900